Exploring the Effect of Mutual Gaze Perception on Students’ Collaborative Discourse

Bertrand Schneider
Stanford University
CS224N final project
schneibe@stanford.edu

ABSTRACT

The goal of this project is to 1) explore a broad set of computational techniques for analyzing educational datasets (i.e. transcripts of students, essays); 2) see if any of those measures sheds a new light on previous results; 3) test whether those metrics have any predictive power regarding learning outcomes. Using the dataset from a previous study, I found that cosine similarity measures are positively correlated with students learning gains when paired with a reference text (such as a textbook chapter). I also explored students’ collaboration, and found that linguistic coordination (i.e. the extent to which students mimic each other in terms of their grammatical structure) did not predict students’ quality of collaboration or learning gains. However, the same measure used at a semantic level (i.e. with word similarity from WordNet) was indeed correlated with those two outcomes. Finally, I used the metrics explored in this paper to feed a machine learning algorithm (SVM) and found that students’ quality of collaboration and learning gains could be “roughly” (using a median split) predicted with an accuracy higher than 90%. Implications for using Natural language processing techniques in education are discussed.

Author Keywords

Natural Language Processing; Eye-tracking; Learning Analytics; Computer-Supported Collaborative Learning.

INTRODUCTION

Despite recent efforts in developing automated ways to analyze students’ discourse [5], most educational researchers still rely on traditional tools to analyze transcripts from students. This includes time-consuming qualitative analyses and manual coding schemes. The field of Natural Language Processing (NLP) has significantly grown and gained in maturity over the past decade, and I suggest that computational techniques can now be advantageously applied to educational datasets. Recent efforts in topic modeling, for instance, seem to be especially promising in terms of gaining insights into students’ discourse and cognitive processes. Unfortunately, social scientists willing to learn those tools are a rare breed, and multi-disciplinary work across education and computer science is slow to appear. For this reason, I propose a first attempt at applying NLP techniques to educational transcripts.

THE CURRENT DATASET

In a previous work [4], I conducted a study on the effect of *mutual visual gaze perception* on students’ collaborative problem-solving processes. In this experiment, dyads of students (groups of two) were asked to remotely collaborate on a set of diagrams to discover how the human brain processes visual information. Each student was in a different room, and could communicate with his/her partner via an audio channel. The information on the screen was similar for both participants. The structure of the activity was as follows: the first step was 12 minutes long; in a second step, students were asked to read a textbook chapter about human vision and discuss their understanding of this topic (12 minutes). Finally, before the analysis activity and after the reading task, students were asked to fill a learning test (pre/post-questionnaires).

Half of our participants were assigned to an experimental group (“visible-gaze”) where they could see the gaze of their partner being displayed in real time on the screen. To achieve this, I used two Tobii X1 eye-trackers running at 30Hz which recorded students’ gaze. In a control group (“no-gaze”), the other half of our participants did not have access to this visualization. I found that this intervention helped students in the first group achieve higher learning gains (Fig. 1) and a higher quality of collaboration (as measured by [2]).

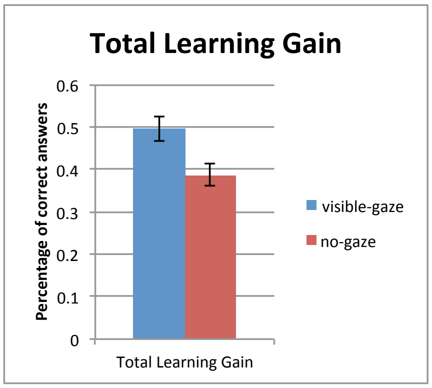

Figure 1: learning gains for the two experimental groups of the study.

Interestingly, by analyzing the eye-tracking data I found that participants in the experimental condition had more moments of joint attention (e.g. they were more likely to be looking at the same diagram at the same time on the screen), and this measure was significantly correlated with positive learning gains. This reinforced the assumption that joint visual attention is a crucial mechanism for coordinating social interactions. In a subsequent qualitative analysis, I also suggested that our intervention helped students because: 1) they were able to anticipate what their partner was about to talk about, because *they could already see where their gaze was*; 2) they could use their gaze as pointers to complement their discourse, and thus remove the need to explicitly mention locations on the diagrams; 3) finally, they could monitor the visual activity of their partner at all time, which helped them establish a common ground.

I propose to use computational techniques to shed a new light on this dataset. More specifically, I would like to explore three aspects of students' dialogues:

1. Are there ways to characterize the effect of my intervention on students' discourse?
2. Is it possible to find markers of productive learning processes?
3. Is it possible to find markers of productive collaborations?

The first question can be answered by designing measures and running statistical tests (i.e. ANOVA) between my two experimental conditions. The second and third questions can be answered by running correlations between my measures of interest, learning gains and collaboration scores.

NATURAL LANGUAGE PROCESSING AND MUTUAL GAZE PERCEPTION

In the next sections, I describe four measures used to provide a preliminary answer to those questions. First, I looked at unigrams, bigrams and trigrams to build categories of interest using a bag of word model. Secondly, I was inspired by information retrieval techniques and decided to perform a tf-idf transformation followed by cosine similarity measures on my transcripts. Thirdly, I looked at the coordination of linguistics styles among my students [1]: in a good collaboration, are students more likely to mimic the grammatical structure of their peers? Fourth, I developed a *coherence* measures to quantify whether or not students were likely to build on their partner's ideas. Finally, I gathered all my measured and ran a machine learning algorithm (Supported Vector Machine) to roughly predict students' learning and quality of collaboration.

Workflow

All my analyses were performed in the IPython Notebook environment, available at the following address[1]. The reader is encouraged to look at the notebook, since I conducted many more analyses than the ones reported here. This work was greatly facilitated by the pandas dataframe library, the scikit learn package, the NLTK functions and corpus, and more generally the enthought distribution (numpy, scipy, matplotlib, among others).

Measure 1: n-grams

To get a sense of my dataset, I first computed unigram, bigram and trigram counts. This helped me understand which words were frequently used and build subsequent categories. For instance, I observed that the word "look" was positively correlated with learning gains (r(37) = 0.42, p = 0.008), which can be associated with either the content to be learnt or a verbal indication to share visual information (e.g. "look at my gaze!"). Overall, it is difficult to interpret frequent n-grams, which is why I grouped them by categories. For instance, the category *anaphora* contained the words "each", "few", "it", "some", "that", "which" and so on; the category *conceptual* discussion contained "think", "cause", "because", "suppose", "impact", and so on. Other categories are shown on the IPython Notebook. Figure two shows the evolution of those two measures during the 12 minutes of the first activity.

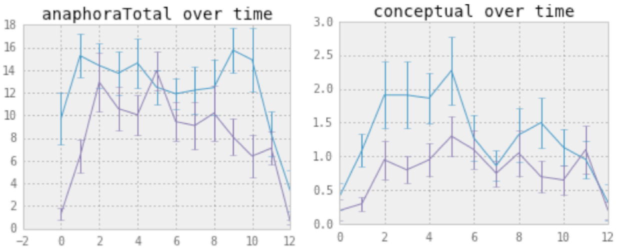

Figure 2: Evolution of anaphoras and words related to conceptual discussion over time. Blue line corresponds to the "visible-gaze" group; purple line corresponds to the "no-gaze" group.

Interestingly, participants in the experimental group used more anaphoras compared to the control group: F(1,41) = 4.88, p = 0.03. This measure can be thought as a proxy for measuring the quality of a common ground between two participants: since anaphoras are ambiguous by nature, they have to be correctly interpreted by the interlocutor and thus indicate a stronger coordination between students. My results suggest that *mutual gaze perception* may be a way to support the establishment of common ground. Additionally, there seem to be some trend showing that more conceptual discussion happened in the "visible-gaze" group (Fig. 2, right side): F(1,41) = 5.52, p = 0.02. One limitation of this measure is that number of words representing this construct is relatively low (between 0 and three words belonging to this category was used every minute).

This first pass on the data is somewhat limited and do noes take full advantage of NLP techniques. In the next section, I use algorithms borrow from the field of information retrieval to find similarities between students.

[1] http://stanford.edu/~schneibe/CS224N/ (please use firefox or safari)

Measure 2: cosine similarity

In this section, I describe how I summarized my data and computed similarity measures to compare students in terms of their quality of collaboration, learning gains and experimental group.

The first step of the process was to apply tf-idf (term frequency–inverse document frequency) to my dataset. Tf-idf is commonly used to summarize a corpus of text. The value of frequent words is increased, but is also offset by their frequency in the corpus; this way, rare words gain a bigger weight and common words (such as "the", "it") gain a smaller weight. This technique is used in information retrieval to score documents' relevance to a query.

We can then compare each students' discourse similarity with other students by using a cosine similarity measure. A cosine similarity measure takes two vectors and computes the angle between them to represent their similarity. Every possible comparison is represented on Figure 2: dark blue lines show students who are very dissimilar to everyone else; hot colors represent similarity. As a sanity check, we can observe that students are identical to themselves (red diagonal); we can also see that groups of students tend to resemble each other (2x2 squares along the diagonal; students in the same group are next to each other on each axis); finally, we can isolate students who are very different from everyone else (e.g. P62 and P63): interestingly, P63 achieved the lowest learning gain on the test.

Additionally, I tried to reorganize students on each axis based on their learning scores (Fig. 4, left side) and their quality of collaboration (Fig.5, right side). The first approach did not cluster students in any meaningful way; however, the second one showed that students with a poor quality of collaboration (left and bottom rows) tend to look very dissimilar to everyone else (shown in dark blue).

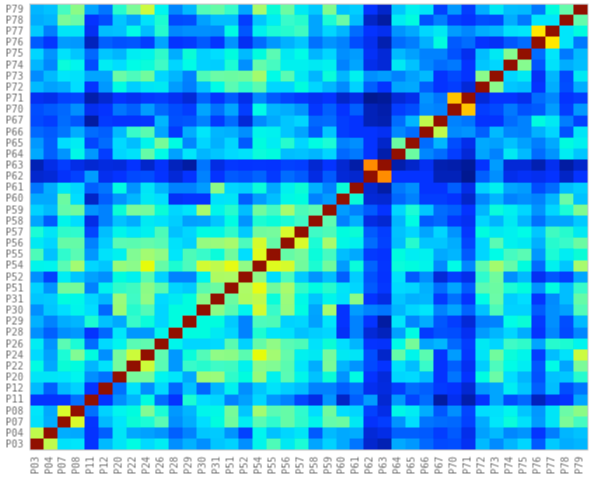

Figure 3: cosine similarity between each participant of the experiment. The diagonal is red, because it represents each students' similarity with herself / himself.

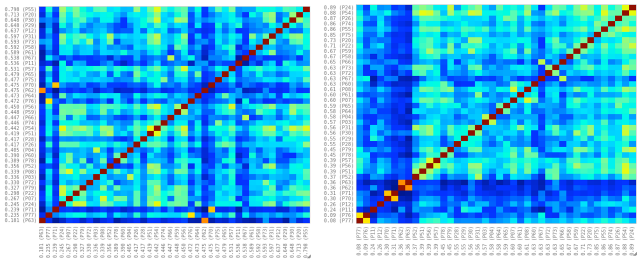

Figure 4: cosine similarity matrix, reorganized with students' learning scores (left side) and quality of collaboration (right side).

In a subsequent step, I tried to look for baselines to compare students with. I used the following two corpora as references: first, I used the best (in terms of her learning score) student of my dataset (P55). She got an impressive 80% gain on the post-test, where the average was around 50%. Second, I inserted the text that students had to read in the experiment into my dataset. This text is highly technical and most likely to pick up students' use of the particular terminology associated with the domain taught.

I found that students in the "visible-gaze" group looked more like P55: F(1,39), p = 0.04, Cohen's d = 0.35 (visible-gaze mean=0.97, SD=0.27; no-gaze mean=0.80, SD=0.20). Interestingly, this measure was positively correlated with students' quality of collaboration: r(38) = 0.545, p < 0.001. There wasn't any difference between the two groups when looking at their similarity with the textbook chapter: F(1,39), p = 0.17, Cohen's d = 0.10 (visible-gaze mean=0.11, SD=0.04; no-gaze mean=0.09, SD=0.04). However, this measure was significantly correlated with students' conceptual understanding of the topic taught: r(38) = 0.335, p = 0.035.

In summary, it looks like taking different baselines is helpful for finding relevant predictors of a good learning group. Taking a student's cosine similarity with an objective reference (i.e. a textbook chapter) seems to be associated with higher learning on a test. Taking a student's cosine similarity with the "best" student of the dataset seems to be associated with productive patterns of collaboration. This makes sense, since students' utterances are representing the way novices discuss a new topic; a scientific text, on the other hand, is produced by experts who master the concepts and terminology of a domain. In sum, those two features could be advantageously used to further explore students' discussion, as well as to feed machine learning algorithms trying to predict students' learning.

Measure 3: Coordination of linguistic styles

Performing tf-idf and cosine similarity measures provides an interesting way to rank students. However it doesn't contribute to our understanding of linguistics patterns used in collaborative learning discussions. To address this issue, I propose to look at the way students build a discourse around the given instructional material. More specifically, I propose to look at a specific phenomenon in social interactions called the *chameleon effect*. In a previous study, Danescu [1] shows how in a social setting people tend to mimic their interlocutor's grammatical structure. Here is an example:

Doc: At least you were outside.

Carol: It doesn't make much difference where you are [...]

From Danescu: "Note that "Carol" used a quantifier, one that is different than the one "Doc" employed. Also, notice that "Carol" could just as well have replied in a way that doesn't include a quantifier, for example, "It doesn't really matter where you are...".

In two large datasets (movie dialogues and twitter), Danescu shows that this effect (called *convergence*) is relatively robust and pervasive. Previous research suggests that this convergence is associated with enhanced communication in organizational contexts and in psychotherapy (cited in [1]). My goal is to 1) replicate Danescu's results on my dataset, and 2) test whether *mutual visual gaze perception* supports convergence.

Concretely, Danescu used 9 categories from the LIWC corpus [3] to compute a converge measures. Those categories are: articles, auxiliary verbs, conjunctions, high-frequency adverbs, impersonal pronouns, negations, personal pronouns, prepositions, and quantifiers. The way convergence is computed is relatively trivial:

$$P(b^t_{\hookrightarrow a} = 1 | a^t = 1) - P(b^t_{\hookrightarrow a} = 1).$$

The first expression is the conditional probability of seeing word type *t* expressed by *b* in answer to *a*, given that *a* used this word type in the previous utterance. The second expression is just the probability of seeing a particular word type in the entire corpus. Subtracting the second expression from the first one gives us a measure of *convergence*.

Figure 5 shows Danescu's result on his dataset. Error bars are shown in red; dark blue bars show the probability of using a particular word type (e.g. articles, pronouns) and light blue bars show the conditional probability of using a particular word type, given that an interlocutor used it in the previous utterance.

Figure 6 shows my replication of Danescu's results. We can see the same pattern emerging: light blue bars (conditional probability that a certain type of words triggers the same word type in the interlocutor's answer) are always higher than the probabilies of this type of word in the corpus. Due to my small corpus, not all differences are statistically significant, but most of them are (i.e. where the standard errors don't overlap).

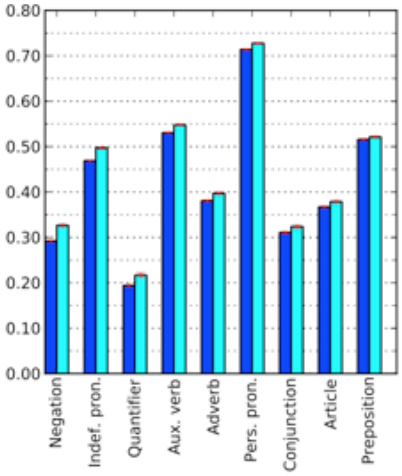

Figure 5: From Danescu [1], this graph shows how people tend to mimic the grammatical structure of their interlocutor. Light blue bars show the conditional probability of using a particular word type, given that an interlocutor used it in the previous utterance. Dark blue bars show the probability of using a particular word type in the entire corpus.

Most importantly, I was interested in using this measure to discriminate between groups of students in my experiment (e.g. "visible-gaze" versus "no-gaze"; productive versus bad collaborators; good versus poor learners). Unfortunately, there wasn't any significant difference between those groups on my convergence measure (F < 1) and no significant correlation. This means that, at least in my corpus, coordination of linguistic styles is not predictive of a good collaboration or positive learning gains. It also shows that *mutual gaze perception* doesn't influence this effect: students are not more likely to imitate each others' grammatical patterns if they can see the gaze of their partner in real time.

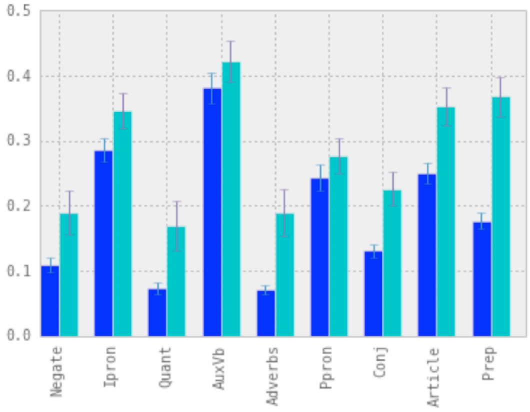

Figure 6: A replication of Danescu's results on my dataset. Errors bars show standard errors. Non-overlapping error bars show statistically significant differences.

This convergence measure, however, only looks at superficial features of collaborative dialogues (i.e. word types). From my point of view, it would be even more interesting to look at the meaning of the word used. If one could show that productive students are more likely to replicate the semantic structure of their partner (i.e. build up on each other's ideas), this would be a much more interesting result.

Measure 4: Coherence

To compute this measure (called *coherence*) in students' discussion, I took advantage of the WordNet service[2]. WordNet is a tree structure connecting each word to its nearest neighbors in terms of its semantic meaning. Various similarity measures have been proposed to compute the similarity of two words: the shortest path between them, the information content of the lowest common subsume of the two nodes (Resnik), and so on. I propose to use this resource to compute meaning similarity between sentences, as opposed to just grammatical structure (as explored above).

The procedure is similar to the one described by Danescu: for each sentence said by A, let's look at the utterance produced by B that immediately precedes it. For each comparison, let's iterate through every word of that sentence A and let's find the most similar word in sentence B. We can then compute the average similarity between those two sentences by keeping the best candidate of each comparison. Repeating this process for each turn produces a vector of scores that we can average to produce a score for each participant.

Additionally, I tried different parameters to explore this space:

1. I looked at the similarity between the current sentence and *n* previous utterances (I tried an n-back of 1, 2 and 3)
2. I used six different similarity measures: shortest Path, Leacock-Chodorow, Wu-Palmer, Jiang-Conrat and Lin.
3. I used two different corpuses: Brown and the semcor datasets
4. I also tried to remove stop words

I found that removing stop words results that were more likely to discriminate between students. The Brown corpus was also more appropriate. However, using different values for the n-back did not change my results (Fig. 7).

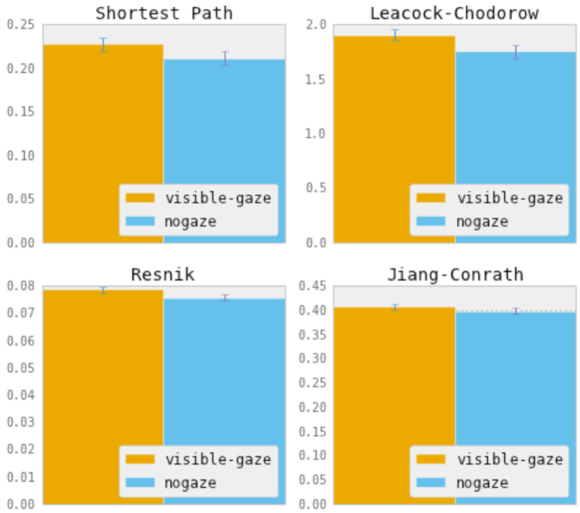

Figure 7: Comparison of four similarity metrics for computing my coherence measure.

Overall, I found that participant in the "visible-gaze" condition were more likely to have a coherent discourse: F(1,39) = 6.90, p = 0.01, Cohen's d = 0.43 (visible-gaze mean=1.97, SD=0.23; no-nogaze mean=1.76, SD=0.25) using the Resnik metric and the Brown corpus. This difference was not significantly different on all combinations of the 4 parameters described above, but the trend was always identical (i.e. students in the visible-gaze group having a higher coherence score). Examples of highly coherent dialogues are shown in table 1.

- So are you talking about lesion two? - Yeah. I'm talking about lesion two.
- It's one, two, three, four. Can you tell me what the forth one down looks like? - The fourth one down looks like the, for the left one, the quarter left side is filled, like the top quarter left side. And for the right one it's the exact same thing.
- I'm thinking lesion five is sort of like the opposite of lesion one. So maybe it's bottom left bottom right? - Well, yeah. OK So then maybe bottom left, yeah, I think that's right.

Table 1. Examples of coherent exchanges.

The examples shown in table 1 shows that my measure seems to pick up exchanges where students use similar words to coordinate themselves: for instance confirming that a student heard a information shared by her partner, or deciding on the next steps of the problem-solving process.

Interestingly, I also found positive correlation between some similarity measures (lin) and dyads' measure of joint

[2] http://wordnet.princeton.edu/

attention: r(18) = 0.574, p = 0.02. This suggests that higher coherence is associated with more gaze recurrence. Some measures were also correlated with dyads' scores on the learning test (e.g. Wu-Palmer): r(18) = 0.522, p = 0.018.

Those results are preliminary, and due to time constraints I didn't have time to further explore them. For instance, I don't believe that I have a perfect understanding of how the differences between those similarity measures impact the coherence score, and how this impact the correlations mentioned above. Additional exploration of highly and poorly coherent sentences would likely help me refine this measure. I am also planning to build my own stop words list, because the list provided by the NLTK package probably removes relevant information from my dataset. Finally, I am interested in refining those results by improving the accuracy of the similarity measure: technical terms like "lateral geniculate nucleus" are not contained in the wordnet dictionary and are central to the learning activity used in my experiment.

In summary, even though those results are interesting, much work still need to be done to understand and improve how coherence develops in small collaborative groups.

Putting the previous results together: predicting students' quality of collaboration and learning gains

The final contribution of this project is to test whether the measures described above have any predictive value. More specifically, can we roughly classify students in terms of their quality of collaboration and learning gains using machine learning algorithms?

To answer this question, I put my hand-labeled categories from section one (measuring n-grams), the cosine similarity scores from section two, the convergence measures from section three and the coherence metrics from section four. The complete dataframe contained 80 features and 40 rows. I used Support Vector Machine (SVM) with a forward search feature selection and tried various kernels (linear, quadratic, polynomial, Gaussian, multilayer perceptron).

For the learning scores, I found that SVM with a multilayer perceptron kernel and 6 features could correctly classify 97.5% of my participants. For the collaboration scores, I found that SVM with a polynomial kernel could correctly classify 92.5% of my participants (table 2).

SVM	**Accuracy**	**Kernel**	**# features**
Learning	97.5%	Multilayer Perceptron	6
Collaboration	92.5%	Polynomial	10

Table 2: Rough classification of students (using a median-split) in terms of their learning gains and quality of collaboration.

Those results are impressive, but they need to be taken with a touch of skepticism. First, I used a lot of features to make this prediction. It is likely that the algorithm is cherry-picking the relevant features to improve its accuracy (which is also over fitting my data). Secondly, my training set is rather small. I only have ~40 students to classify, which is another serious limitation. Finally, I did not use a *held-out* test set; even though I'm using a Leave-One-Out Cross Validation procedure (LOOCV), it's likely that my results are slightly inflated.

In sum, there seems to be some promises in using linguistic features to predict students' learning and ability to collaborate with their peers, but those results need to be replicated on larger datasets with a refined set of features.

DISCUSSION

The goal of this project was to explore various NLP techniques to make sense of educational datasets; I favored a "breadth" approach where I tried promising approaches rather than exploring one specific measure in depth. In future work, I will go back my most promising results (e.g. coherence and cosine similarity) to refine those measures.

To summarize, in this project I found that: 1) n-grams probabilities can help characterize groups of students in terms of building a common ground with their partners (anaphoras); 2) cosine similarity measures are most useful when used with a "reference" corpus (e.g. textbook chapter, transcript of a very good students); 3) coordination of linguistic style has little predictive power in terms of explaining students collaborative learning processes; 4) coherence measures, on the other hand, are associated with those two outcomes; 5) using SVM and the features mentioned above, we can roughly predict students' learning outcomes and quality of collaboration with an accuracy higher than 90%.

Limitations of this work have been highlighted in previous sections (small dataset, limited amount of error analysis). Replicating those results on larger datasets would make a more convincing argument for using NLP measures in education.

CONCLUSION

Overall, this project showed me that NLP approaches hold promises for understanding educational datasets. The measures I described above could easily be applied to other settings, such as forums or online discussion. Future work includes refining those measures and getting a better sense of their predictive value; replicate those results on other datasets; and explore additional topics in NLP (e.g. topic modeling with LSA / LDA).

ACKNOWLEDGMENTS

I would like to thank the CS224N staff for this amazing class and an exciting quarter learning about NLP!

REFERENCES

1. Danescu-Niculescu-Mizil, Cristian, and Lillian Lee. "Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs." *arXiv preprint arXiv:1106.3077* (2011).
2. Meier, Anne, Hans Spada, and Nikol Rummel. "A rating scheme for assessing the quality of computer-supported collaboration processes." *International Journal of Computer-Supported Collaborative Learning* 2.1 (2007): 63-86.
3. Pennebaker, James W., Martha E. Francis, and Roger J. Booth. "Linguistic inquiry and word count: LIWC 2001." *Mahway: Lawrence Erlbaum Associates*(2001): 71.
4. Schneider, Bertrand, and Roy Pea. "Real-time mutual gaze perception enhances collaborative learning and collaboration quality." *International Journal of Computer-Supported Collaborative Learning* 8.4 (2013): 375-397.
5. Sherin, Bruce. "Using computational methods to discover student science conceptions in interview data." *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*. ACM, 2012